Competitive online quantile regression

Raisa Dzhamtyrova[1] and Yuri Kalnishkan[1,2]

[1] Computer Science Department, Royal Holloway, University of London
[2] Laboratory of Advanced Combinatorics and Network Applications, Moscow Institute of Physics and Technology

Abstract. Interval prediction often provides more useful information compared to a simple point forecast. For example, in renewable energy forecasting, while the initial focus has been on deterministic predictions, the uncertainty observed in energy generation raises an interest in producing probabilistic forecasts. One aims to provide prediction intervals so that outcomes lie in the interval with a given probability. Therefore, the problem of estimating the quantiles of a variable arises. The contribution of our paper is two-fold. First, we propose to apply the framework of prediction with expert advice for the prediction of quantiles. Second, we propose a new competitive online algorithm Weak Aggregating Algorithm for Quantile Regression (WAAQR) and prove a theoretical bound on the cumulative loss of the proposed strategy. The theoretical bound ensures that WAAQR is asymptotically as good as any quantile regression. In addition, we provide an empirical survey where we apply both methods to the problem of probability forecasting of wind and solar powers and show that they provide good results compared to other predictive models.

Keywords: prediction with expert advice · online learning · sequential prediction · Weak Aggregating Algorithm · quantile regression · probabilistic forecasting.

1 Introduction

Probabilistic forecasting attracts an increasing attention in sports, finance, weather and energy fields. While an initial focus has been on deterministic forecasting, probabilistic prediction provides a more useful information which is essential for optimal planning and management in these fields. Probabilistic forecasts serve to quantify the uncertainty in a prediction, and they are an essential ingredient of optimal decision making ([4]). An overview of the state of the art methods and scoring rules in probabilistic forecasting can be found in [4]. Quantile regression is one of the methods which models a quantile of the response variable conditional on the explanatory variables ([6]).

Due to its ability to provide interval predictions, quantile regression found its niche in the renewable energy forecasting area. Wind power is one of the fastest growing renewable energy sources ([3]). As there is no efficient way to store wind power, producing accurate wind power forecasts are essential for reliable operation of wind turbines. Due to the uncertainty in wind power generation, there have been studies for improving the reliability of power forecasts to ensure the balance between supply and demand at electricity market. Quantile regression has been extensively used to produce wind

power quantile forecasts, using a variety of explanatory variables such as wind speed, temperature and atmospheric pressure ([7]).

The Global Energy Forecasting Competition 2014 showed that combining predictions of several regressors can produce better results compared to a single model. It is shown in [9] that a voted ensemble of several quantile predictors could produce good results in probabilistic solar and wind power forecasting. In [1] the analogue ensemble technique is applied for prediction of solar power which slightly outperforms the quantile regression model.

In this paper we apply a different approach to combine predictions of several models based on the method of online prediction with expert advice. Contrary to batch mode, where the algorithm is trained on training set and gives predictions on test set, in online setting we learn as soon as new observations become available. One may wonder why not to use predictions of only one best expert from the beginning and ignore predictions of others. First, sometimes we cannot have enough data to identify the best expert from the start. Second, good performance in the past does not necessary lead to a good performance in the future. In addition, previous research shows that combining predictions of multiple regressors often produce better results compared to a single model ([11]).

We consider the adversarial setting, where no stochastic assumptions are made about the data generating process. Our approach is based on Weak Aggregating Algorithm (WAA) which was first introduced in [5]. The WAA works as follows: we assign initial weights to experts and at each step the weights of experts are updated according to their performance. The approach is similar to the Bayesian method, where the prediction is the average over all models based on the likelihood of the available data. The WAA gives a guarantee ensuring that the learner's loss is as small as best expert's loss up to an additive term of the form $C\sqrt{T}$, where T is the number of steps and C is some constant. It is possible to apply WAA to combine predictions of an infinite pool of experts. In [8] WAA was applied to the multi-period, distribution-free perishable inventory problem, and it was shown that the asymptotic average performance of the proposed method was as good as any time-dependent stocking rule up to an additive term of the form $C\sqrt{T}\ln T$.

The WAA was proposed as an alternative to the Aggregating Algorithm (AA), which was first introduced in [12]. The AA gives a guarantee ensuring that the learner's loss is as small as best expert's loss up to a constant in case of finitely many experts. The AA provides better theoretical guarantees, however it works with mixable loss functions, and it is not applicable in our task. An interesting application of the method of prediction with expert advice for the Brier loss function in forecasting of football outcomes can be found in [14]; it was shown that the proposed strategy that follows AA is as good as any bookmaker. Aggregating Algorithm for Regression (AAR) which competes with any expert from an infinite pool of linear regressions under the square loss was proposed in [13].

The contribution of our paper is two-fold. First, as a proof of concept, we apply WAA to a finite pool of experts to show that this method is applicable for this problem. As our experts we pick several models that provide quantile forecasts and then combine their predictions using WAA. To the best of our knowledge prediction with expert advice was not applied before for the prediction of quantiles. Second, we pro-

pose a new competitive online algorithm Weak Aggregating Algorithm for Quantile Regression (WAAQR), which is as good as any quantile regression up to an additive term of the form $C\sqrt{T}\ln T$. For this purpose, we apply WAA to an infinite pool of quantile regressions. While the bound for the finite case can be straightforwardly applied to finite or countable sets of experts, every case of a continuous pool needs to be dealt with separately. We listed above a few results for different specific pools of experts, however there is no generic procedure for deriving a theoretical bound for the cumulative loss of the algorithm. WAAQR can be implemented by using Markov chain Monte Carlo (MCMC) method in a way which is similar to the algorithm introduced in [15], where AAR was applied to generalised linear regression class of function for making a prediction in a fixed interval. We derive a theoretical bound on the cumulative loss of our algorithm which is approximate (in the number of MCMC steps). MCMC is only a method for evaluating the integral and it can be replaced by a different numerical method. Theoretical convergence of the Metropolis-Hastings method in this case follows from Theorems 1 and 3 in [10]. Estimating the convergence speed is more difficult. With the experiments provided we show that by tuning parameters online, our algorithm moves fast to the area of high values of the probability function and gives a good approximation of the prediction.

We apply both methods to the problem of probabilistic forecasting of wind and solar power. Experimental results show a good performance of both methods. WAA applied to a finite set of models performs close or better than the retrospectively best model, whereas WAAQR outperforms the best quantile regression model that was trained on the historical data.

2 Framework

In the framework of prediction with expert advice we need to specify a *game* which contains three components: a space of outcomes Ω, a decision space Γ, and a loss function $\lambda : \Omega \times \Gamma \to \mathbb{R}$. We consider a game with the space of outcomes $\Omega = [A, B]$ and decision space $\Gamma = \mathbb{R}$, and as a loss function we take the pinball loss for $q \in (0, 1)$

$$\lambda(y, \gamma) = \begin{cases} q(y - \gamma), & \text{if } y \geq \gamma \\ (1 - q)(\gamma - y), & \text{if } y < \gamma \end{cases}. \tag{1}$$

This loss function is appropriate for quantile regression because on average it is minimized by the q-th quantile. Namely, if Y is a real-valued random variable with a cumulative distribution function $F_Y(x) = \Pr(Y \leq x)$, then the expectation $\mathbb{E}\lambda(Y, \gamma)$ is minimized by $\gamma = \inf\{x : F_Y(x) \geq q\}$ (see Section 1.3 in [6] for a discussion).

In many tasks predicted outcomes are bounded. For example, wind and solar power cannot reach infinity. Therefore, it is possible to have a sensible estimate for the outcome space Ω based on the historical information.

Learner works according to the following protocol:

Protocol 1

for $t = 1, 2, \ldots$

nature announces signal $x_t \subseteq \mathbb{R}^n$
learner outputs prediction $\gamma_t \in \Gamma$
nature announces outcome $y_t \in \Omega$
learner suffers loss $\lambda(y_t, \gamma_t)$
end for

The cumulative loss of the learner at the step T is:

$$L_T := \sum_{\substack{t=1,\dots,T: \\ y_t<\gamma_t}} (1-q)|y_t - \gamma_t| + \sum_{\substack{t=1,\dots,T: \\ y_t>\gamma_t}} q|y_t - \gamma_t|. \tag{2}$$

We want to find a strategy which is capable of competing in terms of cumulative loss with all prediction strategies $\mathcal{E}_\theta$, $\theta \in \mathbb{R}^n$ (called *experts*) from a given pool, which output $\xi_t(\theta)$ at step t. In a finite case we denote experts $\mathcal{E}_i,\ i = 1, \dots, N$.

Let us denote L_T^θ the cumulative loss of expert $\mathcal{E}_\theta$ at the step T:

$$L_T^\theta := \sum_{\substack{t=1,\dots,T: \\ y_t<\xi_t(\theta)}} (1-q)|y_t - \xi_t(\theta)| + \sum_{\substack{t=1,\dots,T: \\ y_t>\xi_t(\theta)}} q|y_t - \xi_t(\theta)|. \tag{3}$$

3 Weak Aggregating Algorithm

In the framework of prediction with expert advice we have access to experts' predictions at each time step and the learner has to make a prediction based on experts' past performance. We use an approach based on the WAA since a pinball loss function $\lambda(y, \gamma)$ is convex in γ. The WAA maintains experts' weights $P_t(d\theta),\ t = 1, \dots, T$. After each step t the WAA updates the weights of the experts according to their losses:

$$P_t(d\theta) = \exp\left(-\frac{cL_{t-1}^\theta}{\sqrt{t}}\right) P_0(d\theta), \tag{4}$$

where $P_0(d\theta)$ is the initial weights of experts and c is a positive parameter.

Experts that suffer large losses will have smaller weights and less influence on futher predictions.

The prediction of WAA is a weighted average of the experts' predictions:

$$\gamma_t = \int_\Theta \xi_t(\theta) P_{t-1}^*(d\theta), \tag{5}$$

where $P_{t-1}^*(d\theta)$ are normalized weights:

$$P_{t-1}^*(d\theta) = \frac{P_{t-1}(d\theta)}{P_{t-1}(\Theta)},$$

where Θ is a *parameter space*, i.e. $\theta \in \Theta$.

In a finite case, an integral in (5) is replaced by a weighted sum of experts' predictions $\xi_t(i),\ i = 1, \dots, N$.

In particular, when there are finitely many experts $\mathcal{E}_i,\ i = 1, \dots, N$ for bounded games the following lemma holds.

Lemma 1. (Lemma 11 in [5]) *For every $L > 0$, every game $\langle \Omega, \Gamma, \lambda \rangle$ such that $|\Omega| < +\infty$ with $\lambda(y, \gamma) \leq L$ for all $y \in \Omega$ and $\gamma \in \Gamma$ and every $N = 1, 2, \ldots$ for every merging strategy for N experts that follows the WAA with initial weights $p_1, p_2, \ldots, p_N \in [0, 1]$ such that $\sum_{i=1}^{N} p_i = 1$ and $c > 0$ the bound*

$$L_T \leq L_T^i + \sqrt{T}\left(\frac{1}{c}\ln\frac{1}{p_i} + cL^2\right),$$

is guaranteed for every $T = 1, 2, \ldots$ and every $i = 1, 2, \ldots, N$.

After taking equal initial weights $p_1 = p_2 = \cdots = p_N = 1/N$ in the WAA, the additive term reduces to $(cL^2 + (\ln N)/c)\sqrt{T}$. When $c = \sqrt{\ln N}/L$, this expression reaches its minimum. The following corollary shows that the WAA allows us to obtain additive terms of the form $C\sqrt{T}$.

Corollary 1. (Corollary 14 in [5]) *Under the conditions of Lemma 1, there is a merging strategy such that the bound*

$$L_T \leq L_T^i + 2L\sqrt{T \ln N}$$

is guaranteed.

Applying Lemma 1 for an infinite number of experts and taking a positive constant $c = 1$, we get the following Lemma.

Lemma 2. (Lemma 2 in [8]) *Let $\lambda(y, \gamma) \leq L$ for all $y \in \Omega$ and $\gamma \in \Gamma$. The WAA guarantees that, for all T*

$$L_T \leq \sqrt{T}\left(-\ln \int_{\Theta} \exp\left(-\frac{L_T^{\theta}}{\sqrt{T}}\right) P_0(d\theta) + L^2\right).$$

4 Theoretical bounds for WAAQR

In this section we formulate the theoretical bounds of our algorithm.

We want to find a strategy which is capable of competing in terms of cumulative loss with all prediction strategies $\mathcal{E}_\theta$, $\theta \in \Theta = \mathbb{R}^n$, which at step t output:

$$\xi_t(\theta) = x_t'\theta, \tag{6}$$

where x_t is a signal at time t. The cumulative loss of expert $\mathcal{E}_\theta$ is defined in (3).

Theorem 1 *Let $a > 0$, $y \in \Omega = [A, B]$ and $\gamma \in \Gamma$. There exists a prediction strategy for Learner such that for every positive integer T, every sequence of outcomes of length T, and every $\theta \in \mathbb{R}^n$ with initial distribution of parameters*

$$P_0(d\theta) = \left(\frac{a}{2}\right)^n e^{-a\|\theta\|_1} d\theta, \tag{7}$$

the cumulative loss L_T of Learner satisfies

$$L_T \leq L_T^{\theta} + \sqrt{T}a\|\theta\|_1 + \sqrt{T}\left(n\ln\left(1 + \frac{\sqrt{T}}{a}\max_{t=1,\ldots,T}\|x_t\|_\infty\right) + (B - A)^2\right).$$

The theorem states that the algorithm predicts as well as the best quantile regression, defined in (6), up to an additive regret of the order $\sqrt{T}\ln T$. The choice of the regularisation parameter a is important as it affects the behaviour of the theoretical bound of our algorithm. Large parameters of regularisation increase the bound by an additive term $\sqrt{T}a\|\theta\|_1$, however the regret term has a smaller growth rate as time increases. As the maximum time T is usually not known in advance, the regularisation parameter a cannot be optimised, and its choice depends on the particular task. We discuss the choice of the parameter a in Section 6.2.

Proof. We consider that outcomes come from the interval $[A, B]$, and it is known in advance. Let us define the truncated expert $\tilde{\mathcal{E}}_\theta$ which at step t outputs:

$$\tilde{\xi}_t(\theta) = \begin{cases} A, & \text{if } x_t'\theta < A \\ x_t'\theta, & \text{if } A \le x_t'\theta \le B \\ B, & \text{if } x_t'\theta > B \end{cases}. \tag{8}$$

Let us denote $\tilde{L}_T^\theta$ the cumulative loss of expert $\tilde{\mathcal{E}}_\theta$ at the step T:

$$\tilde{L}_T^\theta := \sum_{t=1}^{T} \lambda(y_t, \tilde{\xi}_t(\theta)). \tag{9}$$

We apply WAA for truncated experts $\tilde{\mathcal{E}}_\theta$. As experts $\tilde{\mathcal{E}}_\theta$ output predictions inside the interval $[A, B]$, and predictions of WAA is a weighted average of experts' predictions (5), then each γ_t lies in the interval $[A, B]$.

We can bound the maximum loss at each time step:

$$L := \max_{y\in[A,B],\ \gamma\in[A,B]} \lambda(y,\gamma) \le (B-A)\max(q, 1-q) \le B - A. \tag{10}$$

Applying Lemma 2 for initial distribution (7) and putting the bound on the loss in (10) we obtain:

$$L_T \le \sqrt{T}\left(-\ln\left(\left(\frac{a}{2}\right)^n \int_{\mathbb{R}^n} e^{-\tilde{J}(\theta)} d\theta\right) + (B-A)^2\right), \tag{11}$$

where

$$\tilde{J}(\theta) := \frac{\tilde{L}_T^\theta}{\sqrt{T}} + a\|\theta\|_1. \tag{12}$$

For all $\theta, \theta_0 \in \mathbb{R}^n$ we have:

$$\begin{aligned} \sum_{\substack{t=1,\dots,T:\\ y_t < x_t'\theta}} |x_t'\theta - y_t| &\le \sum_{\substack{t=1,\dots,T:\\ y_t < x_t'\theta}} |x_t'\theta_0 - y_t| + \sum_{\substack{t=1,\dots,T:\\ y_t < x_t'\theta}} |x_t'\theta - x_t'\theta_0| \\ &\le \sum_{\substack{t=1,\dots,T:\\ y_t < x_t'\theta}} |x_t'\theta_0 - y_t| + \sum_{\substack{t=1,\dots,T:\\ y_t < x_t'\theta}} \max_{t=1,\dots,T} \|x_t\|_\infty \|\theta - \theta_0\|_1 \\ &\le \sum_{\substack{t=1,\dots,T:\\ y_t < x_t'\theta}} |x_t'\theta_0 - y_t| + T \max_{t=1,\dots,T} \|x_t\|_\infty \|\theta - \theta_0\|_1. \end{aligned} \tag{13}$$

Analogously, we have:

$$\sum_{\substack{t=1,\dots,T:\\ y_t>x_t'\theta}} |x_t'\theta - y_t| \le \sum_{\substack{t=1,\dots,T:\\ y_t>x_t'\theta}} |x_t'\theta_0 - y_t| + T \max_{t=1,\dots,T} \|x_t\|_\infty \|\theta - \theta_0\|_1. \tag{14}$$

By multiplying inequality (13) by $(1-q)$, inequality (14) by q and summing them, we have:

$$L_T^\theta \le L_T^{\theta_0} + T \max_{t=1,\dots,T} \|x_t\|_\infty \|\theta - \theta_0\|_1. \tag{15}$$

The cumulative loss of truncated expert $\tilde{\mathcal{E}}_\theta$ cannot exceed the cumulative loss of non-truncated expert $\mathcal{E}_\theta$ for all $\theta \in \mathbb{R}^n$:

$$\tilde{L}_T^\theta \le L_T^\theta.$$

By dividing (15) by $\sqrt{T}$ and adding $a\|\theta\|_1$ to both parts, we have:

$$\begin{aligned}\tilde{J}(\theta) \le J(\theta) &\le J(\theta_0) + \sqrt{T} \max_{t=1,\dots,T} \|x_t\|_\infty \|\theta - \theta_0\|_1 + a(\|\theta\|_1 - \|\theta_0\|_1) \\ &\le J(\theta_0) + (\sqrt{T} \max_{t=1,\dots,T} \|x_t\|_\infty + a)\|\theta - \theta_0\|_1,\end{aligned}$$

where

$$J(\theta) := \frac{L_T^\theta}{\sqrt{T}} + a\|\theta\|_1.$$

Let us denote $b_T = \sqrt{T}\max_{t=1,\dots,T}\|x_t\|_\infty + a$. We evaluate the integral:

$$\begin{aligned}\int_{\mathbb{R}^n} e^{-\tilde{J}(\theta)} d\theta &\ge \int_{\mathbb{R}^n} e^{-(J(\theta_0) + b_T\|\theta-\theta_0\|_1)} d\theta \\ &= e^{-J(\theta_0)} \int_{\mathbb{R}} \dots \int_{\mathbb{R}} e^{-b_T \sum_{i=1}^n |\theta_i - \theta_{i,0}|} d\theta_i \\ &= e^{-J(\theta_0)} \int_{\mathbb{R}} \dots \int_{\mathbb{R}} \prod_{i=1}^n e^{-b_T|\theta_i - \theta_{i,0}|} d\theta_i \\ &= e^{-J(\theta_0)} \prod_{i=1}^n \int_{\mathbb{R}} e^{-b_T|\theta_i - \theta_{i,0}|} d\theta_i = e^{-J(\theta_0)} \left(\frac{2}{b_T}\right)^n.\end{aligned}$$

By putting this expression in (11) we obtain the theoretical bound.

Note that even though we apply WAA for truncated experts (8), we achieve the theoretical bound for prediction strategy that competes with a class of experts (6).

5 Prediction Strategy

A prediction of WAA (5) can be re-written as follows:

$$\gamma_T = \int_\Theta \tilde{\xi}_T(\theta) w_{T-1}^*(\theta) d\theta, \tag{16}$$

where

$$w_T^*(\theta) = Z w_T(\theta) = Z \exp\Big(-\frac{1}{\sqrt{T}}\Big(\sum_{\substack{t=1,\dots,T:\\ y_t<\tilde{\xi}_t(\theta)}} (1-q)|y_t - \tilde{\xi}_t(\theta)| + \sum_{\substack{t=1,\dots,T:\\ y_t>\tilde{\xi}_t(\theta)}} q|y_t - \tilde{\xi}_t(\theta)|\Big) - a\|\theta\|_1\Big). \quad (17)$$

and Z is the normalising constant ensuring that $\int_\Theta w_T^*(\theta)d\theta = 1$.

Integral (16) is a Bayesian mixture, where function $\xi_T(\theta)$ needs to be integrated with respect to the normalized distribution $w_T^*(\theta)$. It is possible to avoid the calculation of normalising constant Z as it is a computationally inefficient operation, and integrate function $\xi_T(\theta)$ from the unnormalized distribution $w_T(\theta)$. In order to calculate the integral (16), it is possible to use MCMC algorithms. A good introduction of MCMC for Machine Learning is in [2].

We will use Metropolis-Hastings algorithm for sampling parameters θ from the posterior distribution $\mathcal{P}$. As a proposal distribution we choose Gaussian distribution $\mathcal{N}(0, \sigma^2)$ with some chosen parameter σ. We start with some initial parameter θ^0 and at each step m we update:

$$\theta^m = \theta^{m-1} + \mathcal{N}(0, \sigma^2),\ m = 1, \dots, M,$$

where M is a maximum number of iterations in MCMC method.

The update parameter θ^m at step m is accepted with probability $\min\left(1, \frac{f_\mathcal{P}(\theta^m)}{f_\mathcal{P}(\theta^{m-1})}\right)$, where $f_\mathcal{P}(\theta)$ is the density function for the distribution $\mathcal{P}$ at point θ. At each step by accepting and rejecting the updates of parameters θ we move closer to the maximum of the density function. At the beginning it is common to use a 'burn-in' stage when the integral is not calculated till we will reach the area of high values of the density function $f_\mathcal{P}$. Thus, we perform integration only from the area with high density of $\mathcal{P}$. Some values of θ are accepted even when the calculated probability is less than 1, it allows the algorithm to move away from local minimum of the density function. Because we are interested only in the ratio of density functions of generated parameters, we can generate new parameters θ from the unnormalized posterior distribution $w_T(\theta)$ and avoid the weights normalization at each step which is more computationally efficient.

At time $t = 0$ the algorithm starts with the initial estimate of the parameters $\theta_0 = 0$. At each iteration $t > 0$ we start with parameter θ_{t-1}^M calculated at the previous step $t-1$. It allows the algorithm to converge faster to the correct location of the main mass of the distribution.

WAAQR

Parameters: number $M > 0$ of MCMC iterations,
standard deviation $\sigma > 0$,
regularization coefficient $a > 0$
initialize $\theta_0^M := 0 \in \mathbb{R}^n$

define $w_0(\theta) := \exp(-a\|\theta\|_1)$
for $t = 1, 2, \dots$ do
  $\gamma_t := 0$
  define $w_t(\theta)$ by (17)
  read $x_t \in \mathbb{R}^n$
  initialize $\theta_t^0 = \theta_{t-1}^M$
  for $m = 1, 2, \dots, M$ do
    $\theta^* := \theta_t^{m-1} + \mathcal{N}(0, \sigma^2 I)$
    flip a coin with success probability
      $\min\left(1, w_{t-1}(\theta^*)/w_{t-1}(\theta_t^{m-1})\right)$
    if success then
      $\theta_t^m := \theta^*$
    else
      $\theta_t^m := \theta_{t-1}^m$
    end if
    $\gamma_t := \gamma_t + \tilde{\xi}_t(\theta_t^m)$
  end for
  output predictions $\gamma_t = \gamma_t/M$
end for

6 Experiments

In this section we apply WAA and WAAQR for prediction of wind and solar power and compare their performance with other predictive models. The data set is downloaded from Open Power System Data which provides free and open data platform for power system modelling. The platform contains hourly measurements of geographically aggregated weather data across Europe and time-series of wind and solar power. Our training data are measurements in Austria from January to December 2015, test set contains data from January to July 2016. [3]

6.1 WAA

We apply WAA for three models: Quantile Regression (QR), Quantile Random Forests (QRF), Gradient Boosting Decision Trees (GBDT). These models were used in GEFCom 2014 energy forecasting competition on the final leaderboard ([9]). In this paper the authors argue that using multiple regressors is often better than using only one, and therefore combine multiple model outputs. They noted that voting was found to be particularly useful for averaging the quantile forecasts of different models.

We propose an alternative approach to combine different models' predictions by using WAA. We work according to Protocol 1: at each step t before seeing outcome y_t, we output our prediction γ_t according to (5). After observing outcome y_t, we update experts' weights according to (4).

[3] The code written in R is available at https://github.com/RaisaDZ/Quantile-Regression.

To build models for wind power forecasting we use wind speed and temperature as explanatory variables. These variables have been extensively used to produce wind power quantile forecasts ([7]). We train three models QR, QRF and GBDT on training data set, and then apply WAA using forecasts of these models on test data set. We start with equal inital weights of each model and then update their weights according to their current performance. We estimate the constant of WAA $c = 0.01$ using information about maximum losses on training set.

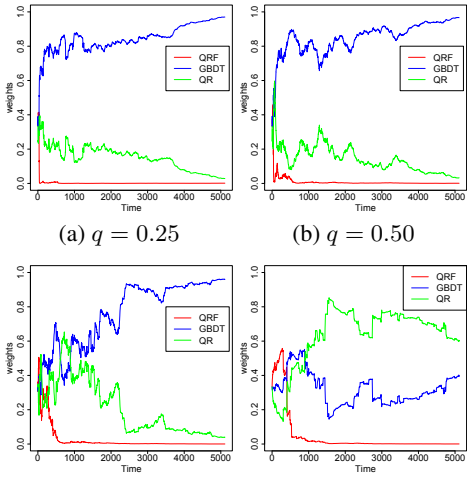
Figure 1 shows weights of each model for different quantiles depending on the current time step. We can see from the graph that for most of quantiles GBDT obtains the largest weights which indicates that it suffers smaller losses compared to other models. However, it changes for $q = 0.95$, where the largest weights are acquired by QR. It shows that sometimes we can not use the past information to evaluate the best model. The retrospectively best model can perform worse in the future as an underlying nature of data generating can change. In addition, different models can perform better on different quantiles.

Table 1 illustrates total losses of QR, QRF, GBDT, WAA and Average methods, where Average is a simple average of QR, QRF and GBDT. For the prediction of wind power, for $q = 0.25$ and $q = 0.50$ the total loss of WAA is slightly higher than the total loss of GBDT, whereas for $q = 0.75$ and $q = 0.95$ WAA has the smallest loss. In most cases, WAA outperforms Average method.

We perform similar experiments for prediction of solar power. We choose measurements of direct and diffuse radiations to be our explanatory variables. In a similar way, QR, QRF and GBDT are trained on training set, and WAA is applied on test data. Figure 2 illustrates weights of models depending on the current step. Opposite to the previous experiments, GBDT has smaller weights compared to other models for $q = 0.25$ and $q = 0.50$. However, for $q = 0.75$ and $q = 0.95$ weights of experts become very close to each other. Therefore, predictions of WAA should become close to Average method. Table 1 shows total losses of the methods. For $q = 0.25$ and $q = 0.5$ both QR and QRF have small losses compared to GBDT, and WAA follows their predictions. However, for $q = 0.75$ and $q = 0.95$ it is not clear which model performs better, and predictions of WAA almost coincide with Average method. It again illustrates that the retrospectively best model could change with time, and one should be cautious about choosing the single retrospectively best model for future forecasts.

Table 1. Total losses ($\times 10^3$)

	wind						solar				
q	QRF	GBDT	QR	Average	WAA	q	QRF	GBDT	QR	Average	WAA
0.25	538.5	**491.2**	516.6	500.3	493.0	0.25	**48.6**	98.3	53.1	63.8	50.1
0.5	757.0	**707.5**	730.7	714.0	709.0	0.5	70.5	110.7	**68.8**	79.1	69.2
0.75	668.3	610.7	633.9	616.6	**610.1**	0.75	63.5	67.6	59.3	58.7	**58.0**
0.95	270.5	222.1	217.5	216.0	**211.0**	0.95	29.2	26.1	23.2	21.0	**20.8**



(a) $q = 0.25$ (b) $q = 0.50$

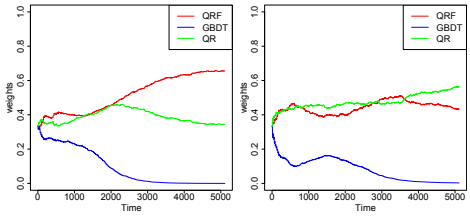
(c) $q = 0.75$ (d) $q = 0.95$

Fig. 1. Weights update for wind power

6.2 WAAQR

In this section we demonstrate the performance of our algorithm for prediction of wind power and compare it with quantile regression model. We train QR on training data set, and apply WAAQR on test set. First, we use training set to choose the parameters of our algorithm. Table 2 illustrates the acceptance ratio of new sampling parameters of our algorithm for $q = 0.5$. Increasing values of σ results in decreasing acceptance ratios of new sampling parameters θ. With large values of σ we move faster to the area of high values of density function while smaller values of σ can lead to more expensive computations as our algorithm would require more iterations to find the optimal parameters. Figure 3 illustrates logarithm of parameters likelihood $w(\theta)$ defined in (17) for $a = 0.1$ and $\sigma = 0.5$ and 3.0. We can see from the graphs that for $\sigma = 3.0$ the algorithm reaches maximum value of log-likelihood after around 800 iterations while for $\sigma = 0.5$ it still tries to find maximum value after 1500 iterations. Table 2 shows the total losses of WAAQR for different parameters a and σ. We can see that choosing the right parameters is very important as it notably affects the performance of WAAQR. It is important to keep track of acceptance ratio of the algorithm, as high acceptance ratio means that we move too slowly and need more iterations and larger 'burn-in' period to find the optimal parameters.

Now we compare performances of our algorithm and QR. We choose the parameters of WAAQR to be the number of iterations $M = 1500$, 'burn-in' stage $M_0 = 300$, regularization parameter $a = 0.1$, and standard deviation $\sigma = 3$. Note that even though we use the prior knowledge to choose the parameters of WAAQR, we start with initial $\theta_0 = 0$ and train our algorithm only on the test set. Figure 4 illustrates a difference between cumulative losses of QR and WAAQR. If the difference is greater than zero, our algorithm shows better results compared to QR. For $q = 0.25$ WAAQR shows



(a) $q = 0.25$ (b) $q = 0.50$

(c) $q = 0.75$ (d) $q = 0.95$

Fig. 2. Weights update for solar power

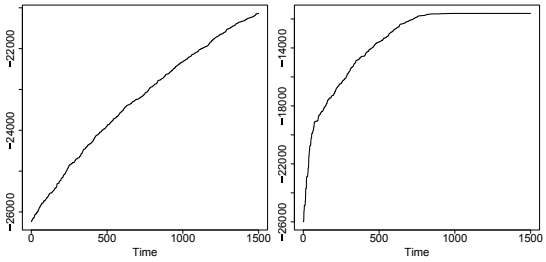
better performance at the beginning, but after around 1000 iterations its performance becomes worse, and by the end of the period cumulative losses of QR and WAAQR are almost the same. We observe a different picture for $q = 0.5$ and $q = 0.75$: most of the time a difference between cumulative losses is positive, which indicates that WAAQR performs better than QR.

Figure 5 shows predictions of WAAQR and QR with $[25\%, 75\%]$ confidence interval for the first and last 100 steps. We can see from the graph, that initially predictions of WAAQR are very different from predictions of QR. However, by the end of period, predictions of both methods become very close to each other.

One of the disadvantages of WAAQR is that it might perform much worse with non-optimal input parameters of regularization a and standard deviation σ. If no prior knowledge is available, one can start with some reasonable values of input parameters and keep track of the acceptance ratio of new generated θ. If the acceptance ratio is too high it might indicate that the algorithm moves too slowly to the area of high values of the probability function of θ, and standard deviation σ should be increased. Another option is to take very large number of steps and larger 'burn-in' period.

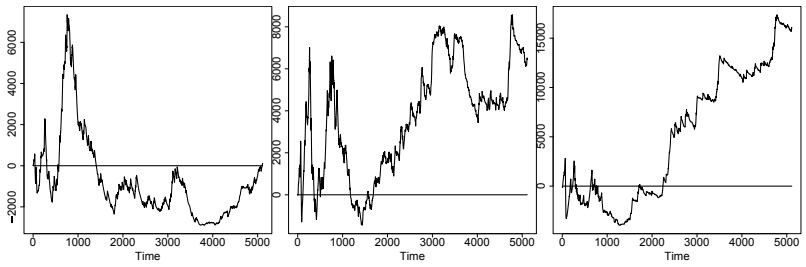
Table 2. Acceptance ratio (AR) and total losses of WAAQR on training set

	AR					Loss			
$a \setminus \sigma$	0.5	1.0	2.0	3.0	$a \setminus \sigma$	0.5	1.0	2.0	3.0
0.1	0.533	0.550	0.482	0.375	0.1	1821.8	823.5	216.3	28.8
0.3	0.554	0.545	0.516	0.371	0.3	1806.2	844.9	265.3	62.7
0.5	0.549	0.542	0.510	0.352	0.5	1815.7	878.5	272.7	92.1
1.0	0.548	0.538	0.502	0.343	1.0	1810.4	877.5	379.3	116.9



(a) $\sigma = 0.5$ (b) $\sigma = 3.0$

Fig. 3. Log-likelihood of parameters for $a = 0.1$.



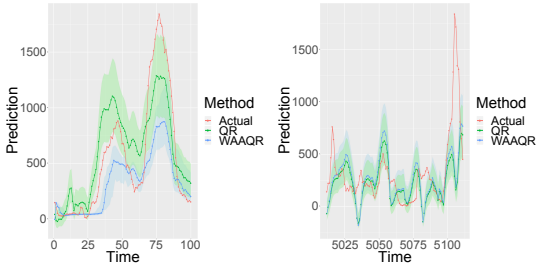
(a) $q = 0.25$ (b) $q = 0.50$ (c) $q = 0.75$

Fig. 4. Cumulative loss difference between QR and WAAQR

7 Conclusions

We proposed two ways of applying the framework of prediction with expert advice to the problem of probabilistic forecasting of renewable energy. The first approach is to apply WAA with a finite number of models and combine their predictions by updating weights of each model online based on their performance. Experimental results show that WAA performs close or better than the best model in terms of cumulative pinball loss function. It also outperforms the simple average of predictions of models. With this approach we show that it is reasonable to apply WAA for the prediction of quantiles.

Second, we propose a new competitive online algorithm WAAQR which combines predictions of an infinite pool of quantile regressions. We derive the theoretical bound which guarantees that WAAQR asymptotically performs as well as any quantile regression up to an additive term of the form $C\sqrt{T}\ln T$. Experimental results show that WAAQR can outperform the best quantile regression model that was trained on the historical data.



(a) first 100 steps (b) last 100 steps

Fig. 5. Predictions with $[25\%, 75\%]$ confidence interval for WAAQR and QR

References

1. S. Alessandrini, L. Delle Monache, S. Sperati, and G. Cervone. An analog ensemble for short-term probabilistic solar power forecast. *Applied Energy*, pages 157: 95–110, 2015.
2. C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan. An introduction to MCMC for machine learning. *Machine Learning Journal*, page 50:5–43, 2003.
3. J. P. Barton and D. G. Infield. Energy storage and its use with intermittent renewable energy. *IEEE Transactions on energy conversion*, pages 19: 441–448, 2004.
4. T. Gneiting and M. Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, pages 1: 125–151, 2014.
5. Y. Kalnishkan and M. Vyugin. The weak aggregating algorithm and weak mixability. *Journal of Computer and System Sciences*, pages 74: 1228–1244, 2008.
6. R. Koenker. Quantile regression. *Cambridge, UK: Cambridge Univ. Press*, 2005.
7. R. Koenker and G. Bassett. Regression quantiles. *Econometrica*, pages 46: 33–50, 1978.
8. T. Levina, Y. Levin, J. McGill, M. Nediak, and V. Vovk. Weak aggregating algorithm for the distribution-free perishable inventory problem. *Operations Research Letters*, pages 38: 516–521, 2010.
9. G. I. Nagya, G. Barta, S. Kazia, G. Borbelyb, and G. Simon. Gefcom2014: Probabilistic solar and wind power forecasting using a generalized additive tree ensemble approach. *International Journal of Forecasting*, pages 32: 1087–1093, 2016.
10. G. O. Roberts and A. F. M. Smith. *Simple conditions for the convergence of the Gibbs sample and Metropolis-Hastings algorithms*. Stoch. Processes Appl., 49, 1993.
11. L. Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, pages 33:1–39, 2010.
12. V. Vovk. Aggregating strategies. In *Proceedings of the 3rd Annual Workshop on Computational Learning Theory*, pages 371–383, San Mateo, CA, 1990. Morgan Kaufmann.
13. V. Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001.
14. V. Vovk and F. Zhdanov. Prediction with expert advice for the Brier game. *Journal of Machine Learning Research*, 10:2445–2471, 2009.
15. F. Zhdanov and V. Vovk. Competitive online generalized linear regression under square loss. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 531–546, 2010.